

AN INTERACTIVE VISUAL ANALYTICS TOOL FOR BIG EARTH OBSERVATION DATA CONTENT ESTIMATION

Daniela FAUR⁽¹⁾, Andreea GRIPARIS⁽¹⁾, Adrian STOICA⁽³⁾, Philippe MOUGNAUD⁽⁴⁾, Mihai DATCU^(1,2)

(1) Politehnica University of Bucharest, Romania

Research Centre for Spatial Information- CEOSpaceTech

(2) German Aerospace Centre, Oberpfaffenhofen, Germany

(3) Terrasigna, Bucharest, Romania

(4) European Space Agency, Esrin, Frascati, Rome

ABSTRACT

This paper introduces a tool designed to provide an innovative and insightful way of exploring Earth observation data content beyond visualization, by addressing a visual analytics process. The considered framework combines machine learning and visualization techniques, empowered through human interaction, to gain knowledge from the data. The proposed tool- eVADE leverages the methodologies developed in the fields of information retrieval, data mining and knowledge representation by the means of a visual analytics component. eVADE increases users capability to understand and extract meaningful semantic clusters together with quantitative measurements, presented in a suggestive visual way.

Index Terms— visual analytics, data visualization, quantitative measurements, Earth observation data content

1. MOTIVATION

The available Earth observation Big Data archives demand insightful resources for information extraction and value adding. Coping with the challenge to make sense of the data, the user may choose the two approaches of the visual data analysis: data visualization and visual analytics, each of them playing a meaningful role in data exploration. A first definition of visual analytics was "the science of analytical reasoning facilitated by interactive human-machine interfaces" [1]. Currently, the visual analytics process targets to couple automated analysis methods with interactive visual representations in order to synthesize information and derive insights from massive, multidimensional amounts of data [2].

There has been an increasing interest in the literature addressing the Visual analytics, imperative in application areas where high dimensional data spaces have to be processed and analyzed. Visualizing the data goes beyond graphical representations, targeting the techniques design to increase the in-

formation entropy of the message [3]. Recent concerns is to develop friendly, highly interactive commercial Visual Analytics tools to support data processing, external database connections and effective data mining algorithms. In the frame of big data analytics approach, aiming to perform landcover classification, the authors of the paper [4] investigate various smart data analytics methods that take advantage of machine learning algorithms and state-of-the-art parallelization approaches in order to overcome limitations of big data processing

An interactive system encompassing the visual analytics approach to explore the comprehensive crime data sets, which contain multidimensional numeric attributes is presented in [5]. Specifically, users could analyze multidimensional data simultaneously in various views and be able to discover high relevant and valuable information through various data combinations.

This paper presents a tool that integrates semi automatic and visual analytics methods in order to derive the statistics of changes for an interest region.

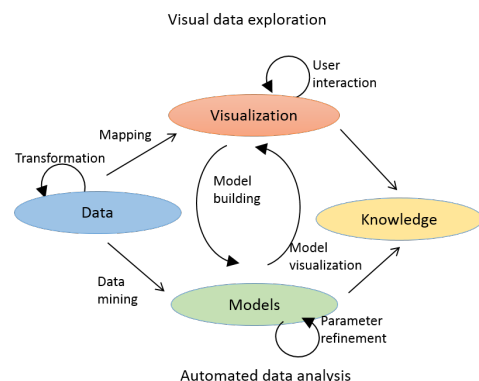


Fig. 1. The visual analytics process as described by Keim in [2]: An abstract overview of the various stages and their transitions.

Thanks to European Space Agency for funding of the eVADE project no.4000120193/17 in the frame of the Romanian Industry Incentive Scheme.

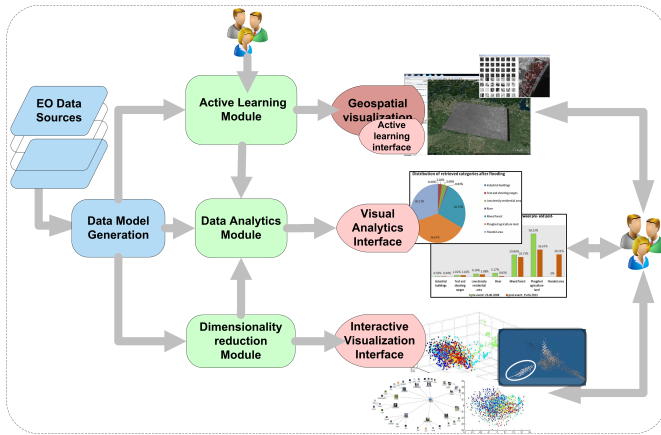


Fig. 2. eVADE tool framework assimilates the generic components of the visual analytics process [1] disclosed in Fig.1.

The first stage is to pre-process and transform the data to derive various representations for further exploration. Further, the analyst is able to select whether to apply visual or automatic analysis methods to initiate the process. If the semi-automated analysis is used first, classical data mining methods are applied to generate models of the data. Once a model is produced, the analyst has to evaluate and refine the model, the interaction with data being an effective way to do this. Visualizations allow the analysts to relate with the automatic methods by modifying parameters or to change the algorithms. Model visualization can then be used to evaluate the findings of the generated models. [2]

2. THE ENGINEERING APPROACH

The design of the eVADE tool - Fig.2 follows the visual analytics process concept (Fig.1) described in [2] through interaction between data, data models, visualization and user perspective in order to gain knowledge about the data.

Data Model Generation assumes the processing of the EO data for: a) extraction of complete metadata, b) tiling the image product in patches, c) estimate for each image patch the relevant descriptors (spectral signatures, texture, structure, etc), d) extract the relevant information from additional information sources (Corine Landcover, Urban Atlas, or in-situ observation, if available).

The **Active Learning Module** module accesses the results of the Data Model Generation (available features, patches and additional information) to support functionalities like query by example, data mining, and semantic labeling of the data content. The active learning algorithm uses a small set of labeled patches, the training data, determined by the user through positive and negative examples selection and the remaining unlabeled patches the test data. The selection

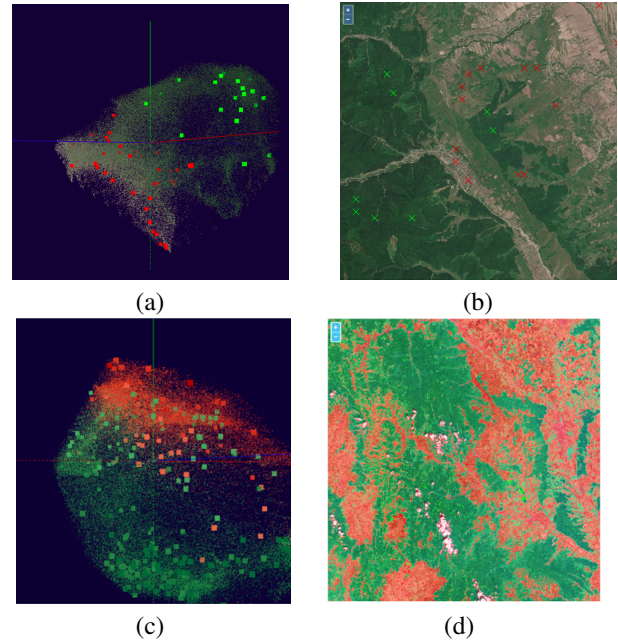


Fig. 3. (a) Geospatial visualization integrated with the Active learning interface-green dots mark the positive samples for the "forest" class while red dots the negative samples;(b) Interactive Visualization Interface displaying the 3D projection of the multidimensional feature space of the scene;the results of user interactions with the tools, labeled scene in geospatial visualization (d) and labeled patches in the 3D projection using for dimensionality reduction t-SNE algorithm [8].

phase is followed by the model learning phase, both stages being alternatively repeated until the user is satisfied with the classification results.

The **Data Analytics Module** ensures the visual analytics frame to clearly understand the data content. This makes it easier to compare datasets and different quantitative findings in a meaningful manner, accessible both to experts and to the untrained users. Data Analytics emphasizes graphical and statistical representations of the data aiming to offer a better understanding of the nature and structure of the data.

The **Dimensionality Reduction Module** aims to perform an appropriate selection of the features (spatial or spectral signatures) and dimensionality reduction algorithms as further inputs for the Data Analytics Module and Interactive Visualization Interface. Dimensionality reduction facilitates classification, visualization and compression of high dimensional data.

eVADE approach will not frame the user into a single visualization tool. The following interfaces will render the EO data, visual and semi automatic derived models and extracted knowledge:

- **Geospatial visualization** integrated with the **Active learning Interface-ALI** - Fig. 3 (b) and (d);

- **Interactive Visualization Interface-IVI**, powered by the **Dimensionality Reduction Module**, exhibits the 3D projection of the multidimensional space of the data - Fig. 3 (a) and (c);
- **Visual Analytics Interface-VAI** providing the graphical means for quantitative, statistical evaluation of the data content - Fig. 6 and Fig. 5.

3. USE CASE SCENARIOS

The main application area is in the field of EO related activities with focus on emergency services. The visual analytics approach that support disaster management adds value to the products by integrating updated and accurate land use/land cover maps in order to delineate and determine the damaged areas. Beyond a simple annotation of the data, the visual analytics frame helps to clearly understand the data content. It becomes easier to compare data sets and various quantitative findings in a coherent manner, attainable for both experts and untrained users. Benefiting from the availability of the Sentinel 2 data, the tool targets to address two major needs: region monitoring for disaster management and land use/land change, deforestation.

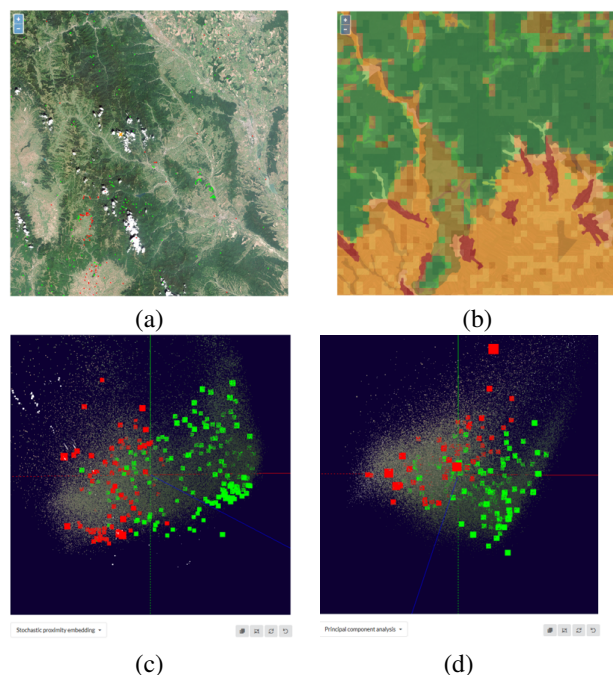


Fig. 4. (a) Sentinel-2 scene revealing a forested area at 04.2018; (b) Corine Landcover classes overlapping the scene, supporting the user in the selection of the positive (green) and negative (red) samples for the training class; the projection of the multidimensional space of the feature vectors using Stochastic proximity embedding algorithm (c) [6] and Principal Component Analysis algorithm (d)[7].

The methodology of data processing depends upon the information a user wants to extract from the data: e.g. in the case of urban areas detection the contextual information, taking into account the neighborhood relationships, is valuable. Hence the need to use the patch level processing. A patch is a small square tile of an image considered as a natural semantic unit of the rendered scene. Each patch will be further described by a feature vector, the extracted information being related to spectral, texture and spatial features of the scene.

The semi-automated analysis of the data begins with the initiation of the active learning process. The user selects an interest region in the 2D geospatial visualization of the scene, by choosing a representative patch for the class that he is interested in, a positive example that will be marked in green; additionally, to strengthen the training procedure, he is able to select patches totally unrepresentative as negative samples”, to be marked in red. The functionality of the tool to alternatively switch between interfaces (IVI and ALI) empowers the user to visualize the selected patches (framed in red or in green) in the 3D projection. That provides means to represent the similarity between patches more directly.

The analyst can select various 3D projections of the multidimensional space in the IVI, achieved through different dimensionality reduction methods. Visualization applied to active learning output ensures great benefits in terms of model interpretation and trust-building. Once the user is satisfied with the results, the obtained class is highlighted both in the 2D and 3D space, and the user begins the labeling of a new class. Fig.3 (a) and (b) show the positive and negative samples in green, respectively in red, given by the user to train the “forest” class, alternatively, in both interfaces.

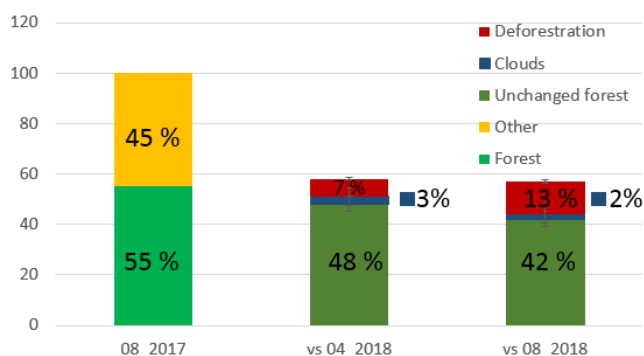


Fig. 5. The statistics of evolution for the “forest” class computed between the scene dated 08.2017 and two different scenes from 04.2018 and 08.2018

The final stage involves the transfer of the labeling results to the Data analytics Modules and their saving in the database; The Visual Analytics interface will display the results in the graphical format chosen by the user, i.e. graphs, charts and plots - Fig. 6 and Fig.5.

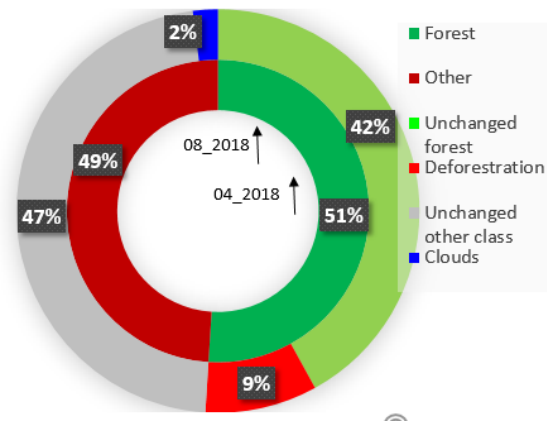


Fig. 6. The statistics of evolution for the "forest" class computed between the scene dated 08.2018 and 04.2018. 51% of the surface of the first scene is forested, a few month later this surface is decreasing 9% percents.

4. RESULTS AND CONCLUSIONS

In order to demonstrate the potential of the tool we address a land use change - deforestation study, considering several Sentinel-2 scene revealing the same region - the Carpathian Mountains in the North-East region of Romania, two years along.

The expected results includes: graphical reports, statistics and situation/semantic maps on: forest area quantitative assessment and statistical analysis related to the evolution of a forested area.

Each scene is split into patches, considering the patches' dimension at 30 pixels. The feature extraction stage aims to describe the content of each patch through the use of specific descriptors. For this study we have used a concatenated feature vector meaning that we have included the spectral signature, texture and the Weber Local Descriptors [7]. The derived multidimensional space was projected in the IVI using three dimensionality reduction methods: Principal Component Analysis, Stochastic Proximity Embedding and t-Distributed Stochastic Neighbor Embedding [6] [7].

Part of the preliminary results of the workflow are represented in Fig. 4, Fig. 6 and Fig.5, emphasizing the potential of this value-added tool able to extend, beyond human perception, the visualization of information content of EO and to statistically deliver this content.

5. REFERENCES

[1] James J. Thomas and Kristin A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, National Visualization and Analytics Ctr, 2005.

[2] Daniel A. Keim, Joern Kohlhammer, G. Elis, and Florian Mansmann, *Mastering the Information Age. Solving Problems with Visual Analytics*, 2010, Eurographics Association, Goslar.

[3] M. Chen and H. Jaenicke, "An information-theoretic framework for visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1206–1215, Nov 2010.

[4] G. Cavallaro, M. Riedel, J. A. Benediktsson, M. Goetz, T. Runarsson, K. Jonasson, and T. Lippert, "Smart data analytics methods for remote sensing applications," in *2014 IEEE Geoscience and Remote Sensing Symposium*, July 2014, pp. 1405–1408.

[5] D. Li, Y. Wang, S. Wu, J. Qi, and T. Wang, "An visual analytics approach to explore criminal patterns based on multidimensional data," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2017, pp. 5563–5566.

[6] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik, "Dimensionality reduction: a comparative," 2009.

[7] A. Griparis, D. Faur, and M. Datcu, "Feature space dimensionality reduction for the optimization of visualization methods," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2015, pp. 1120–1123.

[8] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[9] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend, "High-resolution global maps of 21st-century forest cover change," *Science*, vol. 342, no. 6160, pp. 850–853, 2013.